June 1974

THE STRUCTURE OF TASK ORIENTED DIALOGS

by

Barbara G. Deutsch
Artificial Intelligence Center

Technical Note 90
SRI Project 1526

Proceedings IEEE Speech Symposium, Carnegie-Mellon University, Pittsburgh, Pennsylvania, April 15-19, 1974.

The work reported herein was sponsored by the Advanced Research Projects Agency of the Department of Defense under Contract DAHC04-72-C-0009 with the U. S. Army Research Office.

Summary

The discourse and task information in task oriented dialogs and their use in a speech understanding system are discussed in this paper. The results of analyzing some task oriented dialogs are given. A preliminary model of the structure of these dialogs and heuristics for building and using it in a speech understanding system are presented.

Introduction

It is often stated that sentences cannot be understood in isolation, and that many of the ambiguities that occur would disappear if the sentence were considered in some context (and that those ambiguities that remain are ambiguities for people as well). People are not, usually, intentionally ambiguous. When a sentence can be interpreted in several ways, there are cues -- either linguistic (things recently talked about) or situational (things both speaker and listener can "see") -- which enable the listener to choose the correct interpretation. This leads us to consider including a discourse history and some model of discourse structure in a language system. The discourse history would give us access to past linguistic events in the order in which they occurred; the discourse structure would give us a way to interpret the information in that history. But first, we must know what information is available from discourse and how it can be used. The context in which we have been studying these problems is that of task oriented dialogs. By a task oriented dialog we mean one in which two (or more) people communicate for the sole purpose of completing some task.

Recently, linguists have been devoting a fair amount of effort to discourse analysis. Grimes[1] contains an extensive reference list. Much of the work has concentrated on looking at text. There has also been work on spoken monologues (such as folktales[2]) and some work on dialogs. Most of this work has focussed on describing various characteristics of discourse. Some of the questions asked are: How does the speaker decide what information to include? How does the expression of new and old information differ? How are different kinds of information -- setting, actors, and events -- conveyed? What techniques are used to make the text cohesive? These questions and their answers are relevant to task oriented dialog also. In addition, task oriented dialog has several characteristics which derive from the close interaction between the speech and the situational context. References are often to objects in the situation. Even when the people conversing are not in the same physical location, to communicate effectively they must share a model of the current state of the world in the task environment. The dialog includes frequent instances of checking the communication channel to make sure it is operating[3]. There is frequent updating of the common model of the state of the world.

Previous work on dialog has centered on describing characteristics of dialog and the parts played by the different speaker/listeners involved. We are interested not only in characterizing task oriented dialogs, but also in finding those features that are amenable to formalization and eventually incorporation in a computer program to understand speech. In this paper, we report some initial observations we have made and some preliminary design ideas.

System Framework

A computer system which would serve as an expert consultant to a human apprentice doing maintainance of small electromechanical devices in a workstation environment is being built at SRI[4]. The workstation domain includes a work table, a toolbox, various tools (such as wrenches, hammers, screwdrivers, and wheelpullers), parts (nuts, bolts, screws), and small devices that need repair. The computer system is expected to be able to give the apprentice advice about how to assemble and disassemble the equipment and how to diagnose and repair faults. To do this, the system will have to be able to describe the parts and tools in the workstation, explain how to use the tools, and answer any (task oriented) questions the apprentice may have. It will also have to be able to understand progress reports from the apprentice so that it can update its world model. The level of detail of advice will depend on the sophistication of the apprentice: more inexperienced apprentices will need more detailed advice.

Natural language will be an important communication channel in this system. The current SRI speech understanding system[5] will be the basis of the natural language component. This system has been using the repair of leaky faucets as the task domain for its initial work. The goal has been to allow a human expert to ask questions and give directions to a (simulated) robot about the faucet world. The planning component of this system is far simpler than what will be required when the computer is the advice giver. The syntax and semantics of the current system will also have to be modified to deal with the more complicated workstation domain. We have run some experiments to get data on the kind of language the speech component in such a system would have to handle.

Simulation Experiments

We have been taping dialogs between two people working to complete repair tasks in the workstation environment. The particular device we have used for our experiments is a small air compressor. One person plays the role of an expert adviser; the other acts as an apprentice.

Our initial experiments were done with the expert and the apprentice in the same room. In some of the experiments the expert and the apprentice could see each other; in other experiments they could not. A major characteristic of these dialogs was a cooperative completion of ideas: one of the dialog participants would start a sentence, and the other would complete it. That is, as soon as the listener thought he knew what the speaker was trying to say, he would indicate this by completing the thought. Looking over these dialogs we discovered that sentence fragments were at least as common as complete sentences. Communication is seen as a subpart of the whole task -- namely, cooperating to get the task done.

For a speech system to be able to understand speech like this, it would have to have an extremely strong semantic component. Before trying to build such a system, we decided to run a series of experiments to see how communication would be affected if the expert and the apprentice were not allowed to interrupt each other.

A second observation from our first set of experiments was that the amount of vision available has a large effect on the language used. When the two participants can see each other, there is a much larger amount of deictic (i.e., pointing) reference. Since our system will only have limited vision capabilities, we also designed the second set of experiments so that visual information was restricted. Thus, this second set of experiments serves as a closer simulation of the system for which we are building a speech component.

The design of the experiment is shown in Figure 1. The apprentice, the

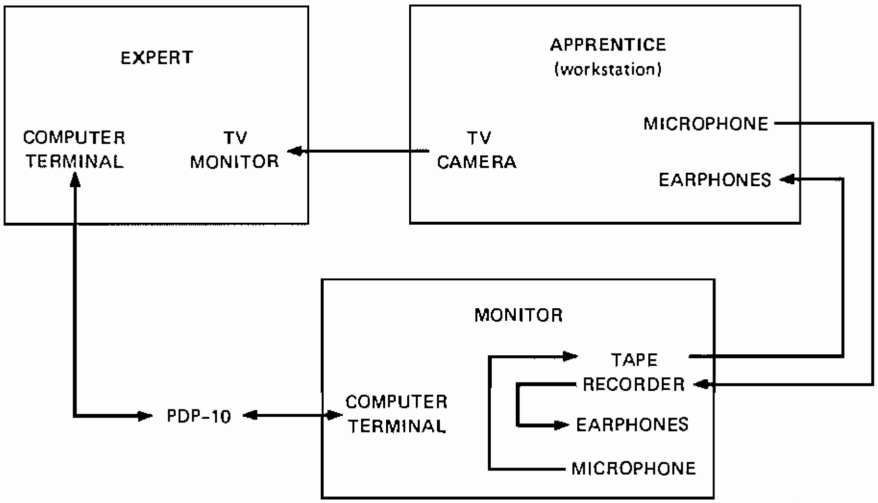


Figure 1

expert, and a monitor who served as a link between them were all in separate rooms. The apprentice and the monitor were connected by microphone and earphone links. The expert and the monitor were linked through computer connections to a PDP-10. The monitor typed what the apprentice said to the expert and read what the expert typed to the apprentice. In addition, the monitor was responsible for seeing that the apprentice did not speak while the expert was typing. Since only one person could type at a time (a feature of the linking program), the expert could not interrupt the apprentice. The whole dialog was taped and a typescript file was kept. When the expert wanted to see something he had to request a picture; only still shots were allowed. A camera operator was used so the apprentice would not be disturbed when the expert asked for a picture of something. This also allowed the apprentice to point at parts and tools when he wanted to identify them for the expert.

General Observations

It would be reasonable to expect that the speech input from someone taking directions would be very limited, consisting mostly of: "yes", "no", and "I don't understand." Our experience indicates this is not the case. The apprentice often takes the initiative in a task. He may need to explain a problem, or ask a question about an instruction he has just been given; he may want to propose the next step, or report on what he has been doing. Answers to questions are often far more complicated than a simple yes or no. The amount and kind of speech vary with the level of expertise of the apprentice. The dialog with a naive apprentice is filled with definitions of terms (establishing a common vocabulary), and questions about how certain operations should be done. With an experienced apprentice there is less talk in general; the dialog consists mostly of reports of what has been done or explanations of what the apprentice intends to do next.

Placing a monitor between the expert and the apprentice had the effect of slowing down the dialog. However the response time did not seem to be much greater than what we can expect in the next few years from a speech understanding system. The main effect of this slowdown seemed to be that the apprentice would go ahead and try things while waiting for a response. Aside from the slowdown caused by messages being typed, the participants did not seem at all hampered by this means of communication. In fact what appears to have happened is that the listener would act on his understanding (before an utterance was completed) rather than finishing the utterance (i.e., the speech act) and then acting.

Discourse and Task Information

The discourse context (history of the dialog) and the situational context (state of the task and workstation environment) provide two kinds of information to the speech understanding system. First, they limit the domain under consideration in interpreting an utterance. In a system with a predictive parser[6] this limitation is useful both in resolving references once an utterance is parsed and in limiting the lexicon which is considered in attempting to parse an utterance. For example, if the apprentice has been using two wrenches recently, it is these and not the others in the workstation he is referring to when he asks, "Do you know where I put the wrenches?"

The second use of task and discourse information is to check the results of other parts of the system for consistency with the current state of the discourse and to furnish information for resolving inconsistencies. For example, "No" is not a sufficient answer to, "Do I clamp the wheelpuller to the rim of the wheel?" This question indicates the apprentice does not know how to use the wheelpuller and wants instructions.

Although task and discourse information are separable, there are important ways in which they interact. For example, a model of the current task situation is crucial to understanding the following piece of dialog:

E:Remove the pulley.
A:Do I have to remove the screws or just loosen them?

There are two setscrews which hold the pulley on its shaft. It is these screws to which the apprentice is referring in his question. But these screws have not been mentioned previously in the dialog. The reference is exophoric (i.e., it is a reference to something in the situation, not the text). Thus, task specific information is needed to provide the situational context in which discourse procedures can resolve reference.

Current Semantics and Pragmatics

In the current SRI speech understanding system, the world model provides the basis for a task model. It consists of assertions about properties of objects that exist in the domain and relations between them, and a set of procedures for operating on these objects. This model will have to be augmented by a detailed model of the devices and parts in the workstation. The procedures for operating on the objects in the domain will have to contain more information about how they interact. Interaction with a sophisticated planning component will be necessary. These additions will enable reasonable predictions about future utterances to be made and help resolve problems like that in the preceding example.

The current system has a limited discourse component. This includes a set of semantic routines for handling anaphoric reference. The antecedent for a third person pronoun is determined by searching the previous utterance. (Note, at present only pronouns parsed in case slots in the main clause of an utterance are handled by the system.) Semantic features of the pronoun and its antecedent must match. Since the pronoun has been parsed in the context of a particular verb's case frame, the antecedent must meet the semantic feature tests for the pronoun's case slot. One goal of analysis of the dialog experiments was to determine whether these routines could resolve the references that occurred in the dialogs. It appears, as we will show later, that with some modification they can.

Dialog Structure

The structure of a task oriented dialog closely parallels the structure of the task being accomplished. For example, where an assembly task involves a succession of steps, the dialog may consist of a sequence of subdialogs of the form:

GET NEXT PART -> PUT IN PLACE -> FASTEN.

Each of these subparts may be many sentences long and may itself contain subparts; for example: FASTEN might include

DETERMINE HOW FASTENED ->
GET TOOLS (& PARTS) -> TIGHTEN.

A reasonable analogy is that looking at the structured history of the dialog is like looking at an outline of the task as performed by the particular apprentice. Some parts of the task are explored in more detail than others. The corresponding part of the "outline" has more detailed levels. Two things at the same level in the "outline" do not necessarily have to follow each other in that order. In fact, this is one of the places where the discourse history differs from a plan for carrying out the task. The discourse history keeps track of the order in which utterances were spoken. It is not concerned with whether this order is a necessary one.

The most interesting property of this hierarchical structure is that references operate mostly within a subpart. So, for example, inside of the actual subdialog for GET TOOL there may be references (e.g., "it", them") to different tools. However, once the tool has been located, references operate between elements of the subdialog for FASTEN. What happens is that once a subtask is completed, and hence that part of the dialog (i.e., that subdialog) exited, the subdialog is effectively removed from the focus of attention. However, it appears to be labelled in some manner so that it can be retrieved and looked at later if necessary. When it does become necessary to refer to something that occurred in that subdialog, it will be retrieved (via its label) and placed back in focus. The label is usually a short statement of what the subtask achieved. For example, consider the following piece of dialog:

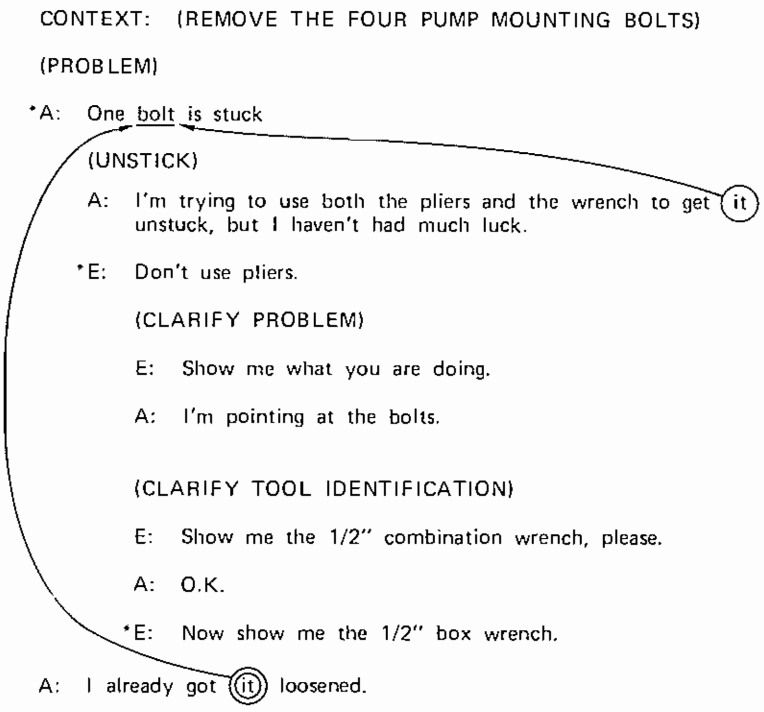
> A:...I'm having trouble getting the wheel off.
> E:Use the wheelpuller. Do you know how to use it?
> A:No.
> E:Do you know what it looks like?
> .
> .
> .
> A:I have the jaws around the hub. How should I take it off now?
> E: Tighten the screw in the center of the puller . . . that should slide the wheel off the shaft.
> A: OK. It's off.
> A: A little metal semicircle fell off when I took the wheel off.

The statement "OK.It's off." closes the subdialog corresponding to the subtask "REMOVE WHEEL". Note that even though the last utterance in this dialog fragment comes immediately after the "REMOVE WHEEL" subdialog has been closed, the subdialog must be retrieved through its label. The apprentice does this by saying "when I took the wheel off."

Subdialogs are not always explicitly closed linguistically as in the preceding example. In those dialogs from the first set of experiments in which vision was used, most of the subdialogs were closed through visual communication rather than linguistically (i.e. both people could see that a subtask was completed). There are also instances when the linguistic closure is made clear by a reference to something that occurred at a higher level in the discourse. For example, consider the following dialog:

A:One bolt is stuck. I'm trying to use both the pliers and the wrench to get it unstuck, but I haven't had much luck.
E: Don't use pliers. Show me what you are doing.
A: I'm pointing at the bolts.
E: Show me the 1/2" combination wrench, please.
A:OK
E: Good, now show me the 1/2" box wrench.
A: I already got it loosened.

"It" in the last utterance refers to "bolt". Looking at the structured history in Figure 2, we see that "bolt" is two levels up in the dialog structure. This pronoun reference implicitly closes the <clarify tool> and <unstick> subdialogs.



CONTEXT: (REMOVE THE FOUR PUMP MOUNTING BOLTS)

(PROBLEM)

*A: One bolt is stuck

(UNSTICK)

A: I'm trying to use both the pliers and the wrench to get it unstuck, but I haven't had much luck.

*E: Don't use pliers.

(CLARIFY PROBLEM)

E: Show me what you are doing.

A: I'm pointing at the bolts.

(CLARIFY TOOL IDENTIFICATION)

E: Show me the 1/2" combination wrench, please.

A: O.K.

*E: Now show me the 1/2" box wrench.

A: I already got it loosened.

*denotes utterances looked at in trying to find an antecedent for it.

Figure 2

For the dialog structure to be useful to a computer program for understanding natural language, it must be possible to detect these implicit closures easily. Otherwise it would be possible to derive the structure only after the dialog was understood -- not a very useful aid to understanding! A careful inspection of how pronoun references are used reveals the following very interesting property. If the antecedent to the pronoun reference does not occur earlier in the sentence, or in the preceding sentence, then it occurs in the last utterance of some higher level in the structure. (We note in passing that forward pronoun references do not occur in any of the dialogs we have collected.) That means we never have to look at more than one utterance at any level to resolve the reference. In the preceding example, looking back linearly, "bolt" is seven utterances back in the dialog history; in the structured dialog history "bolt" occurs in the last utterance two levels up. The anaphoric routines only have to be modified to look up the discourse structure until a suitable antecedent is found. In our example, the objects mentioned in the other two candidate last utterances (marked by * in the figure), namely pliers and wrench, are not "loosenable" and thus do not satisfy the case frame requirements for the antecedent to "it".

There are other problems in maintaining the discourse structure. Interaction with the planning component of the system will be crucial in determining subtasks and possible problems at a given stage in the task, and in labelling the subdialogs. At present, we envision the planning component of the system producing plans in the form of an acyclic graph that encodes a partial ordering of the steps of the plan. Operations that must occur in a certain order are distinguished from those which can be accomplished in any order. For example a partial plan to remove the pump is shown in Figure 3. The arrows show the partial ordering. An operation at the head of an arrow cannot be done until the operation at the tail of that arrow is done. This representation helps both in predicting likely future utterances (i.e., limiting the context) and in constructing the discourse structure. Assume the last utterance expressed the goal represented by some point in the graph. The next utterance

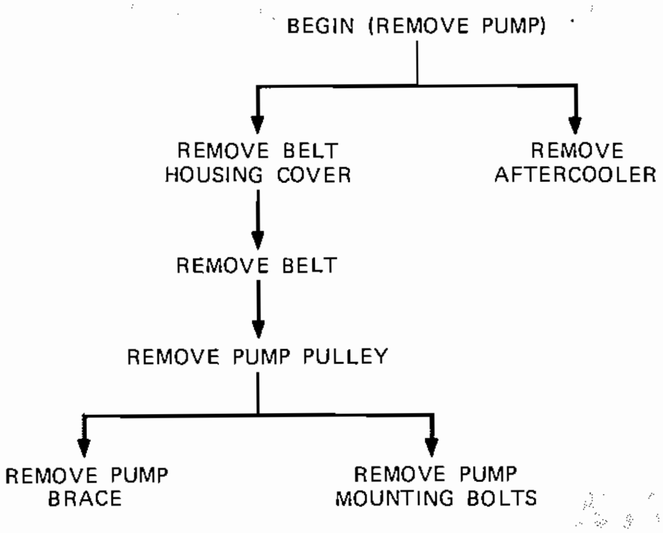


Figure 3

can be one of three types: (1)it can give or ask for more information about that goal; (2) it can express completion of that goal; or (3) it can express the goal represented by some other point in the graph. In the example of Figure 3, if the apprentice has just been told to remove the pump pulley, likely next utterances are:

(1)How do I do it? or, Do I have to pull the screws all the way out?
(2)OK, it's off.
(3)Why can't I take the aftercooler off first?

It is clear how this aids the discourse structure. An utterance of type (1) adds a new entry to the structure at a lower level. An utterance of type (2) or (3) causes a new entry at the same level.

Work also has to be done on mapping linguistic statements into general operations and relations. This is necessary, for example, to understand that "got it loosened" solves the problem "bolt is stuck".

Conclusions

Since communication in this highly interactive sense is a major function of language, task oriented dialog seems to be a promising area for studying the syntactic and semantic devices people use to communicate effectively. Looking at task oriented dialogs reveals several characteristics that can be used to aid a natural language understanding system. We have concentrated on exploring one, namely, the structure of the dialog and its relation to the task. This structure is useful in resolving reference and in limiting the context considered at any point in the dialog. We have outlined possible heuristics to be incorporated in an understanding system to take advantage of this structure. It is clear that further analysis of the dialogs would be beneficial both in revealing other discourse characteristics and in further determining the syntax and semantics used in such dialogs.

References

1. J. Grimes, "The thread of discourse," Technical report no. 1, Cornell University Department of Modern Languages and Linguistics, Ithaca,N.Y., 1972.
2. W.L. Chafe, Contrastive semantics project, First technical report, U.C. Department of Linguistics, Berkeley, CA., 1972.
3. J.J. Robinson, "Performance grammars," Proc. IEEE Symposium on Speech Recognition (to appear)
4. N.J. Nilsson, et al, "Plan for a computer-based consultant system," SRI Artificial Intelligence Center working paper, Jan. 1974.
5. D.E. Walker, "The SRI speech understanding system," Proc. IEEE Symposium on Speech Recognition (to appear)
6. W.H. Paxton, "A best first parser," Proc. IEEE Symposium on Speech Recognition (to appear)